STA 260s20 Assignment Four: Confidence Intervals Part Two[1]

These homework problems are not to be handed in. They are preparation for Quiz 4 (Week of Feb. 10) and Term Test 2. **Please try each question before looking at the solution**.

1. Here are some distribution facts that are helpful to know without looking at a formula sheet. You are responsible for the proofs of these facts too, but here you are just supposed to write down the answers.

    (a) Let $X \sim N(\mu, \sigma^2)$ and $Y = aX + b$, where a and b are constants. What is the distribution of Y?

    (b) Let $X \sim N(\mu, \sigma^2)$ and $Z = \frac{X-\mu}{\sigma}$. What is the distribution of Z?

    (c) Let $X_1, \ldots, X_n$ be a random sample from a $N(\mu, \sigma^2)$ distribution. What is the distribution of $Y = \sum_{i=1}^{n} X_i$?

    (d) Let $X_1, \ldots, X_n$ be a random sample from a $N(\mu, \sigma^2)$ distribution. What is the distribution of the sample mean $\overline{X}$?

    (e) Let $X_1, \ldots, X_n$ be a random sample from a $N(\mu, \sigma^2)$ distribution. What is the distribution of $Z = \frac{\sqrt{n}(\overline{X}-\mu)}{\sigma}$?

    (f) Let $X_1, \ldots, X_n$ be independent random variables, with $X_i \sim N(\mu_i, \sigma_i^2)$. Let $a_1, \ldots, a_n$ be constants. What is the distribution of $Y = \sum_{i=1}^{n} a_i X_i$?

2. Use the formula sheet as necessary for this question.

    (a) Let $X_1, \ldots, X_n$ be independent random variables with $X_i \sim \chi^2(\nu_i)$ for $i = 1, \ldots, n$. Find the distribution of $Y = \sum_{i=1}^{n} X_i$. Show your work. Your answer includes a statement of the parameter(s).

    (b) Let $X_1, \ldots, X_n$ be random sample from a $N(\mu, \sigma^2)$ distribution. Using your answers to earlier questions in this assignment, find the distribution of $Y = \frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \mu)^2$. Your answer includes a statement of the parameter value(s).

    (c) Let $Y = Y_1 + Y_2$, where Y_1 and Y_2 are independent, $Y_1 \sim \chi^2(\nu_1)$ and $Y \sim \chi^2(\nu_1 + \nu_2)$, where ν_1 and ν_2 are both positive. Derive the distribution of Y_2. Your answer includes a statement of the parameter value(s).

3. Let $X_1, \ldots, X_n \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$. The sample variance is $S^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}$.

    (a) The difference $X_j - \overline{X}_n$ is a linear combination of the form $\sum_{i=1}^{n} a_i X_i$, as in Question (1f). What are the coefficients $a_1, \ldots, a_n$?

    (b) Show $Cov(\overline{X}, X_j - \overline{X}) = 0$ for every $j = 1, \ldots, n$.

    (c) How do you know that $\overline{X}$ and s^2 are independent? You may use without proof the fact that for the normal distribution, zero covariance implies independence, and that functions of independent random variables are also independent.

    (d) Show that $\sum_{i=1}^{n} (X_i - \mu)^2 = \sum_{i=1}^{n} (X_i - \overline{X})^2 + n (\overline{X} - \mu)^2$.

    (e) Prove that $Y_1 = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$

---

4. Again, let $X_1, \ldots, X_n$ be random sample from a $N(\mu, \sigma^2)$ distribution.

(a) The t distribution is defined as follows. Let $Z \sim N(0,1)$ and $Y \sim \chi^2(\nu)$, with Z and Y independent. Then $T = \frac{Z}{\sqrt{Y/\nu}}$ is said to have a t distribution with ν degrees of freedom, and we write $T \sim t(\nu)$. Using results from earlier questions, prove $T = \frac{\sqrt{n}(\overline{X} - \mu)}{S} \sim t(n-1)$. Be sure to indicate why your Z and Y are independent.

(b) You can see that T is a "pivotal quantity." It's a random variable that is a function of the parameter, but whose distribution does not depend on the parameter value. Starting with a probability statement about the pivotal quantity, derive an exact $(1 - \alpha)100\%$ confidence interval for μ. "Derive" means show all the high school algebra. Your answer is a pair of formulas, one for the lower confidence limit and one for the upper confidence limit.

(c) The t distribution was introduced by William Gossett, writing under the name Student. The reference is Student (1908). "The probable error of a mean," *Biometrika* 6, 1-25. Gossett illustrated the method using two measurements on ten patients suffering from insomnia (trouble sleeping). Each number is a difference, representing how much *extra* sleep the patient got when taking a sleeping pill, compared to a baseline measurement. Drug 1 is Dextro-hyoscyamine hydrobomide, while Drug 2 is Laevo-hyoscyamine hydrobomide. Each patent tried both drugs, with a recovery period of several days between trials. Here are the data:

Patient	Drug 1	Drug 2
1	0.7	1.9
2	-1.6	0.8
3	-0.2	1.1
4	-1.2	0.1
5	-0.1	-0.1
6	3.4	4.4
7	3.7	5.5
8	0.8	1.6
9	0.0	4.6
10	2.0	3.4

The strategy here is to compute a difference for each patient, Drug 1 minus Drug 2. The difference represents how much more sleep the patient got when using Drug 1. The differences are $X_1, \ldots, X_{10}$, and μ is the expected advantage of Drug 1 over Drug 2. Give a point estimate and a 95% confidence interval for μ. The point estimate is a single number. The confidence interval is a pair of numbers, the lower confidence limit and the upper confidence limit.

(d) Does the confidence interval allow you to decide which drug worked better?

(e) Derive an exact $(1 - \alpha)100\%$ confidence interval for σ^2. "Derive" means show all the high school algebra. Your answer is a pair of formulas, one for the lower confidence limit and one for the upper confidence limit. You can locate the pivotal quantity in your answers to earlier questions.

(f) Using the data from Question (4c), give a 95% confidence interval for σ^2. The answer is a pair of numbers, the lower confidence limit and the upper confidence limit.

5. It is natural to be interested in the difference between two expected values. For example, volunteer patients in a clinical trial might be randomly assigned to receive either blood pressure medicine A or blood pressure medicine B. One way to ask which medicine is more effective is to compare the two expected blood pressures. Or, we might wonder whether there is a difference between men and women in their expected score on a racism questionnaire.

Accordingly, let $X_1, \ldots, X_{n_1} \overset{i.i.d.}{\sim} N(\mu_1, \sigma^2)$, and $Y_1, \ldots, Y_{n_2} \overset{i.i.d.}{\sim} N(\mu_2, \sigma^2)$. In addition, the X_i are independent of the Y_j. This is a model for random sampling from two separate populations. Notice that while the two expected values might be different, the two normal distributions in this model have the same variance. This is for technical convenience, as you will see.

(a) The quantity of primary interest here is $\mu_1 - \mu_2$. Give a natural point estimator of $\mu_1 - \mu_2$.

(b) What is the distribution of $\overline{X} - \overline{Y}$? You should be able to just write it down.

(c) Standardize the difference to obtain Z. This will be the numerator of T.

(d) Using the fact that the sum of two independent chi-squares is chi-squared, propose a nice Y variable to go in the denominator of T. How do you know that Z and Y are independent?

(e) Write a formula for your T random variable, and simplify. The cancellation of σ^2 in numerator and denominator is wonderful. This is why the variances of the two populations are assumed equal; it's not to make the model more realistic.

(f) The formula for T is usually written

$$T = \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \quad \text{where } S_p = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}$$

The statistic S_p is called the "pooled" estimated standard deviation. Simplify your T formula a bit more until it has this form.

(g) What are the degrees of freedom of T? That is, what is the parameter ν?

(h) Derive an exact $(1 - \alpha)100\%$ confidence interval for $\mu_1 - \mu_2$. Your answer is a pair of formulas, one for the lower confidence limit and one for the upper confidence limit.

(i) Two surgeons in a cosmetic surgery practice decide to have a friendly competition. The wait list has 20 patients who want surgery to make their noses smaller. Ten patients are randomly assigned to Surgeon A, and the other ten are assigned to Surgeon B. A panel of medical students rate the facial appearance of the patients on a 100 point scale before surgery and again six weeks after. The number for each patient is improvement (according to the medical students): After minus before. Of course, the medical students are not told which doctor did the surgery.

Because of scheduling problems and drop-out (people change their minds), Surgeon A only did nine surgeries, and Surgeon B did seven. So with $n_1 = 9$ and $n_2 = 7$, we have $\overline{x} = 14.1$, $s_1^2 = 48.2$, $\overline{y} = 13.3$, $s_1^2 = 32.7$. Give a point estimate and a 95% confidence interval for $\mu_1 - \mu_2$. The point estimate is a single number, and the confidence interval is a pair of numbers, a lower confidence limit and an upper confidence limit.

(j) Who wins the contest?

6. The F distribution is defined as follows. Let $Y_1 \sim \chi^2(\nu_1)$ and $Y_2 \sim \chi^2(\nu_2)$ be independent. Then $F = \frac{Y_1/\nu_2}{Y_2/\nu_2}$ is said to have an F distribution with ν_1 and ν_2 degrees of freedom, and we write $F \sim F(\nu_1, \nu_2)$. For obvious reasons, ν_1 is sometimes called the "numerator degrees of freedom," and ν_2 is called the "denominator degrees of freedom."

   (a) What is the support of the F distribution?

   (b) Show that if $F_1 \sim F(\nu_1, \nu_2)$, then
      i. $F_2 = 1/F_1 \sim F(\nu_2, \nu_1)$.
      ii. $P(F_1 \leq x) = P(F_2 \geq 1/x)$.

   (c) For a single random sample from a normal distribution, you have shown that $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$. So for two independent random samples, it's easy to locate a pivotal quantity for the ratio $\frac{\sigma_1^2}{\sigma_2^2}$ based on the F distribution. Write the pivotal quantity. What are the degrees of freedom?

   (d) Derive a 95% confidence interval for $\frac{\sigma_1^2}{\sigma_2^2}$.

   (e) Using the statistics from Question (5i), give a 95% confidence interval for $\frac{\sigma_1^2}{\sigma_2^2}$. The tough part is the critical values. The F tables in the back of the text (not part of the formula sheet) have some explanation up front. Your answer is a pair of numbers, a lower confidence limit and an upper confidence limit.

   (f) The confidence interval for the difference between means depends on the assumption that $\sigma_1^2 = \sigma_2^2$. Does your confidence interval for $\frac{\sigma_1^2}{\sigma_2^2}$ suggest that this assumption is incorrect?